University of Houston-Downtown

64/100 = 60

Name: Arzaka

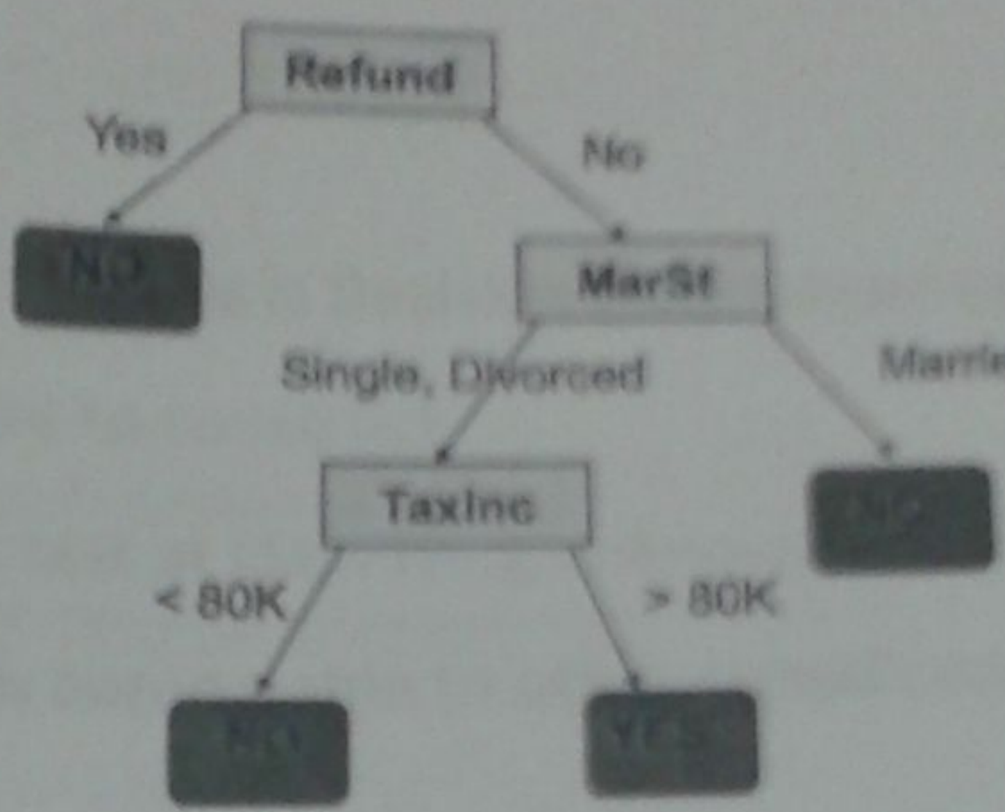
ID: 9002409

Date: 03/14

CS4319 – Data Mining & Warehouses
Midterm 2

1. Decision Trees are examples of:
   a. Clustering Techniques
   b. Classification Techniques
   c. Association Rule Mining Techniques
   d. Regression Tree Techniques

2. Consider the following decision tree.



Refund	Marital Status
No	Married

What would be the answer for the record above (replace the ? wit

a. Yes

b. No

c. Can not be decided and decision tree should be grown further

d. Can not be decided and decision tree should be pruned

3. Training Set → Learn Model is also called _____
   Apply Model → Test Set is also called _____

a. Deduction - Induction
b. Induction - Deduction
c. Learning-Validating
d. Testing-Validating

UHD
University of Houston-Downtown

4. Which of the following is (are) Measures of Node Impurity?

I. Gini Index
II. Entropy
III. Misclassification error

a. I
b. I & II
c. III only
d. I, II, and III

5. GINI question

(C)

6. Which of the following is Decision Tree Advantages:

I. Inexpensive to construct

II. Extremely fast at classifying unknown records

III. Accuracy is comparable to other classification techniques for many simple da

a. I
b. I & II
c. III only
d. I, II, and III

7. Given two models of similar generalization errors, one should prefer the simpl the more complex model, is the definition of

a. Simple Model Theory
b. Occam's Razor
c. Basic Model Principle
d. None of the Above

8. The method where you reserve 2/3 for training and 1/3 for testing is

a. Holdout
b. Cross Validation
c. Bootstrap
d. Stratified Training

UHD
University of Houston-Downtown

9.

Cost Matrix	PREDICTED CLASS		
	C(i|j)	+	-
ACTUAL CLASS	+	-1	100
	-	1	0

Model M_1	PREDICTED CLASS		
		+	-
ACTUAL CLASS	+	150	40
	-	60	250

Model M_2	PREDICTED CLASS		
		+	-
ACTUAL CLASS	+	250	45
	-	5	200

Accuracy of M1: 80 %

Accuracy of M2: 90 %

Cost (M1): 3(+)

Cost (M2): 5(+)

10. Given two models of classification:

- Model M1: accuracy = 85%, tested on 30 instances
- Model M2: accuracy = 75%, tested on 5000 instances

What test would help to find which model is better?

a. Test of Accuracy
b. Test of Reliability
c. Test of Significance
d. Test of Comparability

11. What are the goals of clustering:

I. Minimizing Intra-cluster distances
II. Maximizing Intra-cluster distances
III. Minimizing Inter-cluster distances
IV. Maximizing Inter-cluster distances

a. I & III
b. II & IV
c. I only
d. IV only

UHD
University of Houston-Downtown

12. K-means is
    a) Medoid-based Hierarchical clustering
    b) Medoid-based Partitional clustering approach
    c) Centroid-based Hierarchical clustering
    (d) Centroid -based Partitional clustering approach

13. Which of the following is true for K-means:
    I. The centroid is typically the mean of the points in cluster
    II. Most of the convergence happens in approximately mid-time of clustering
    III. Complexity is $O(n^2)$

    (a) I & III
    b. II only
    c. I only
    d. I, II, and III

14. Which one is the most common measure to evaluate K-Means Clusters:

    a. Cohesion
    b. Separation
    (c.) SSE
    d. Cluster Mean

15. Which of the following is solution to Initial Centroids Problem in K-means:
    I. Multiple Runs
    II. Bisecting K-means
    III. Sample and use hierarchical clustering to determine initial centroid
    IV. Post Processing
    V. Using k2 centroids instead of k in initialization

    a. I, II, IV
    (b.) I, II, V
    c. I, II, III, IV
    d. I, II, III, IV, V

16. Which one is not one of the limitations of K-Means?

    a. K-means has problems when the data contains outliers
    b. K-means has problems with differing cluster sizes
    c. K-means has problems with non-globular shapes
    (d.) K-means has problems with large number of clusters

University of Houston-Downtown

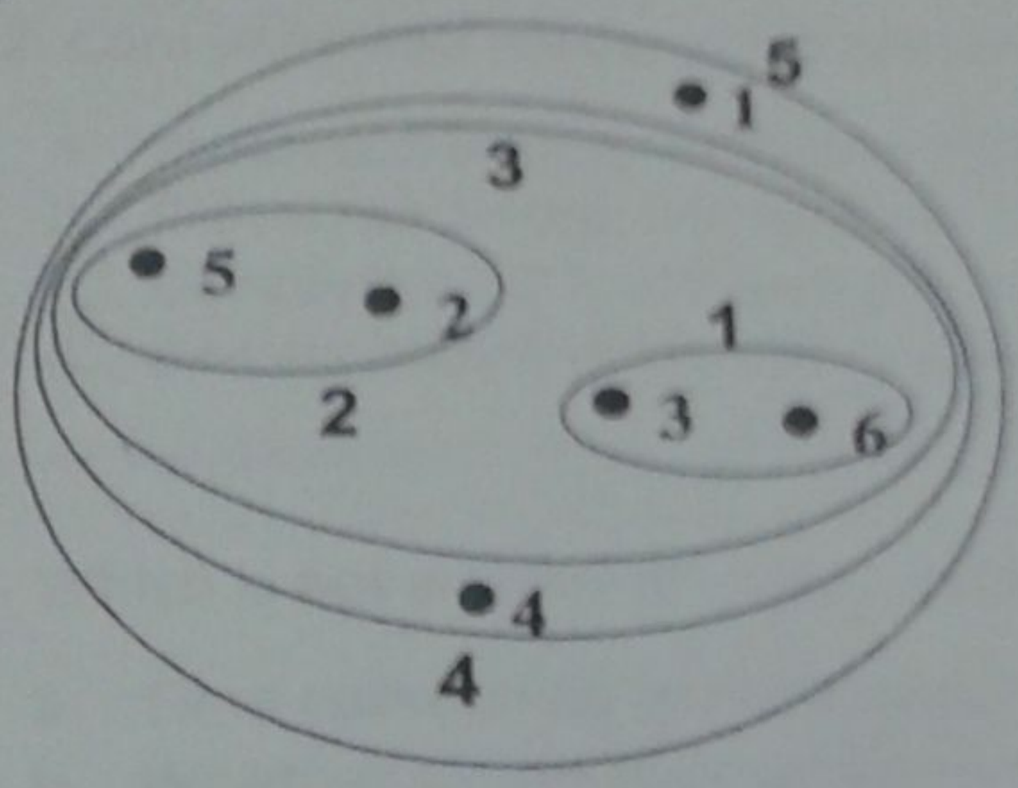
17. Using _____ for Cluster Similarity has following disadvantages:

- Biased towards globular clusters
- Tends to break large clusters

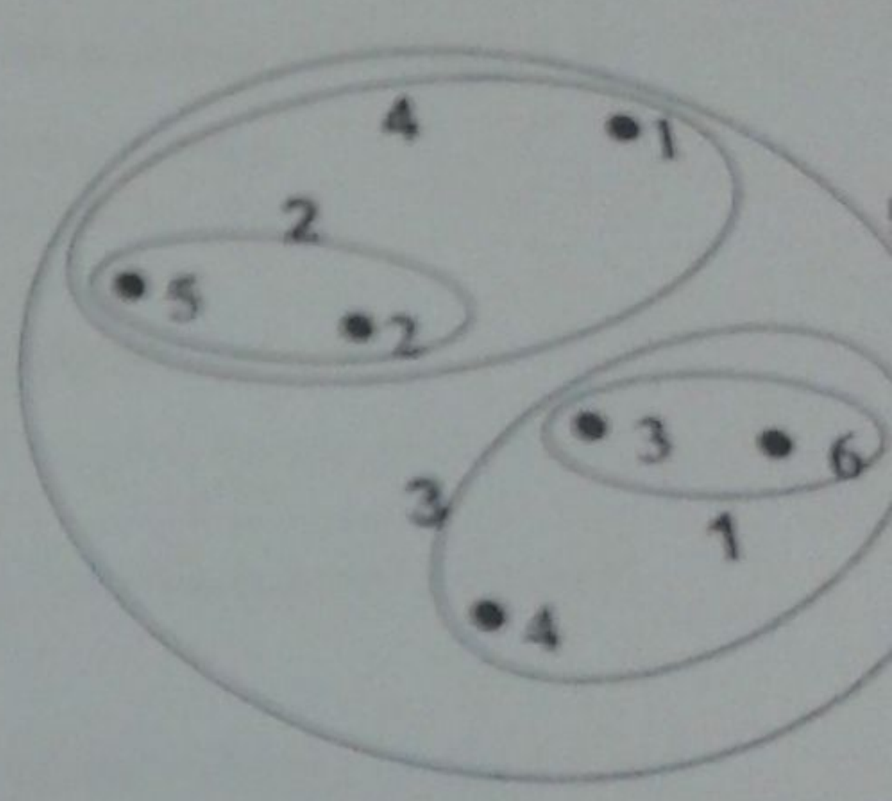
Which of the following should be inserted in _____?

a. MIN
b. MAX
c. Group Average
d. Objective Function

18. Fill out the blanks with MIN, MAX, Group Average, Ward's



MIN



MAX

19.

Entropy is a _____ type of index:

a. Internal Index
b. External Index
c. Relative Index
d. Silhouette Index

20.

Entropy is a _____ type of index:

a. Internal Index
b. External Index
c. Relative Index
d. Silhouette Index

_____ measures how closely related are objects in a cluster

_____ measures how distinct or well-separated a cluster is from other clusters

a. Cluster Cohesion - Cluster Separation
b. Cluster Separation - Cluster Cohesion
c. Cluster Similarity - Cluster Distance
d. Cluster Distance – Cluster Similarity

2. Core point and border point are concepts in an algorithm that is a

a. Prototype-based, Partitional Clustering Algorithm
b. Prototype-based, Hierarchical Clustering Algorithm
c. Density-based Partitional Clustering Algorithm
d. Density-based Hierarchical Clustering Algorithm

3. Which is not a different when comparing K-means and DBSCAN:

a. Time complexity of O(m) whereas
b. Whether or not they produce same set of cluster at each run,
c. Whether or not using all attributes in finding clusters
d. Whether or not they are supervised or unsupervised

4. The method that predicts a value of a given continuous valued variable based on the her variables, assuming a linear or nonlinear model of dependency is :

a. Correlation
b. Regression
c. Cohesion
d. Separation

5. Measures that satisfy all 3 conditions of {positivity, symmetry, and Triangle Inequa lled:

a. Objective measures
b. Metrics
c. Distance Function
d. Similarity Function